Modèle de prédiction du concours d'Economie

Guillaume CORRE

Table des matières

1 Question 1

1.1 Etude statistique

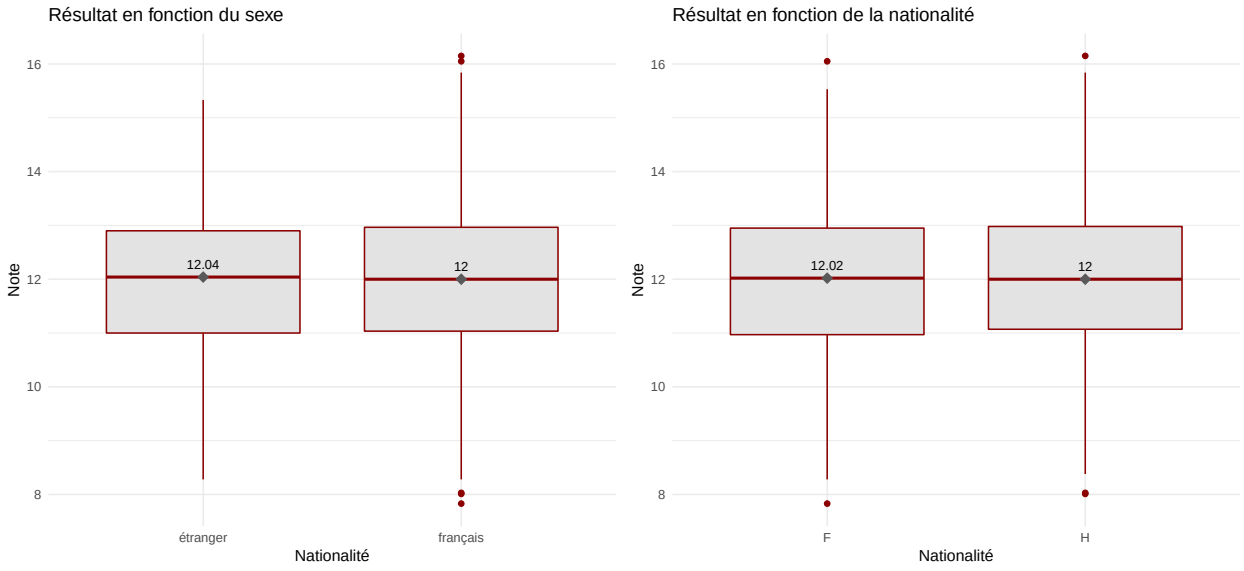


TABLE 1 – Taux de réussite en fonction du sexe et de la nationalité

	Sexe		Nationalité	
	F	**H**	**étranger**	**français**
non	0.4930925	0.4955401	0.488	0.495873
oui	0.5069075	0.5044599	0.512	0.504127

On remarque que les variables sexe et nationalité ne semblent pas discriminer la note, et donc ne discrimine pas l'admissibilité. On peut dire que le fait d'être une femme ou un homme ne change pas le taux de réussite, tout comme le fait d'être étranger ou non. Grâce au tableau, on remarque que le taux est aux alentours de 50% de réussite et 50% d'échec pour chaque modalités de chaque variables. Ces deux variables ne sont donc pas discriminantes.

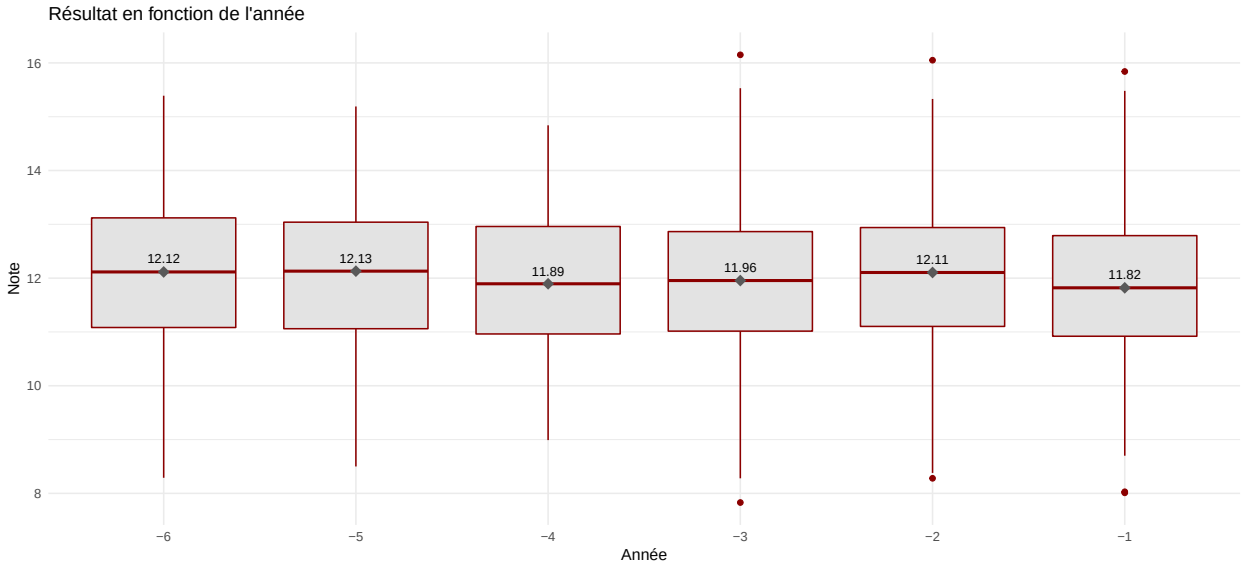


Table 2 – Taux de réussite en fonction de l'année

	Année					
	-6	**-5**	**-4**	**-3**	**-2**	**-1**
non	0.4645161	0.4478114	0.5228758	0.5115607	0.4763314	0.5354108
oui	0.5354839	0.5521886	0.4771242	0.4884393	0.5236686	0.4645892

L'année ne semble également avoir un impact majeur sur le taux d'admissibilité. En effet, lorsque l'on observe le tableau, l'on voit que le taux de réussite est toujours entre 0.46 et 0.55 entre les différentes années. Cette variable ne semble pas discriminer. Le taux de réussite est constant entre les années, nous n'utiliserons pas cette variable dans nos futurs modèles.

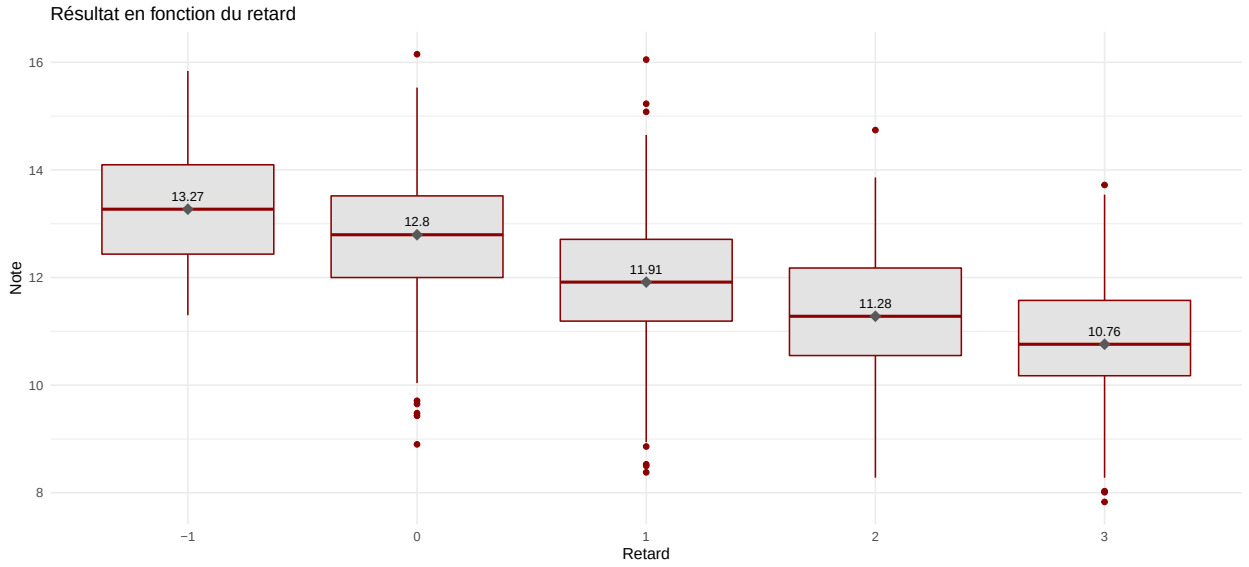
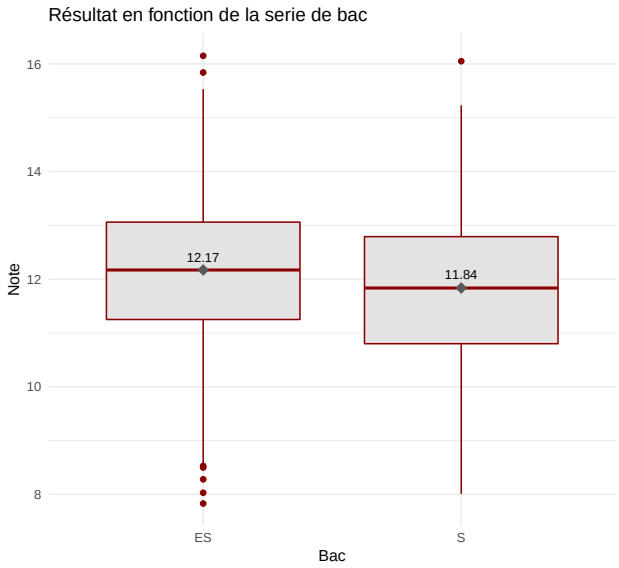


Table 3 – Taux de réussite en fonction du retard

	Retard				
	-1	**0**	**1**	**2**	**3**
non	0.1445783	0.2492918	0.5251938	0.6917293	0.8469657
oui	0.8554217	0.7507082	0.4748062	0.3082707	0.1530343

Le retard lui, semble avoir un impact très important sur la note obtenu par les candidats et donc sur l'admissibilité. Plus le retard est important, plus la note est mauvaise. On voit clairement sur le tableau que le taux de réussite passe de 85% pour un élève en avance à 15% pour un élève avec 3 ans de retard. Cette variable discrimine.



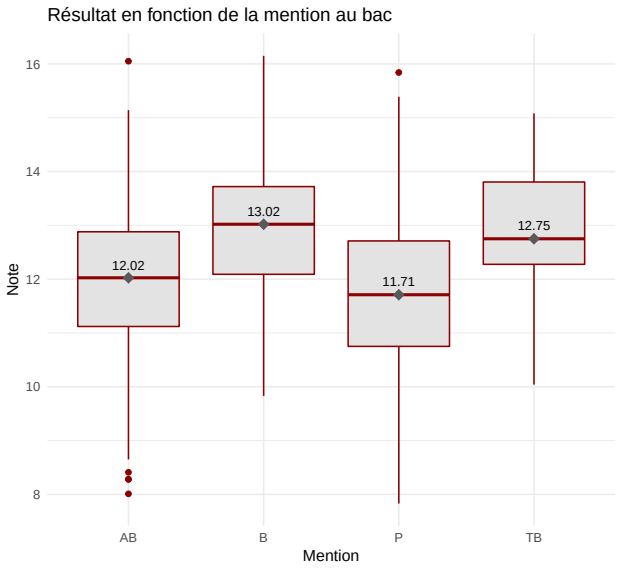
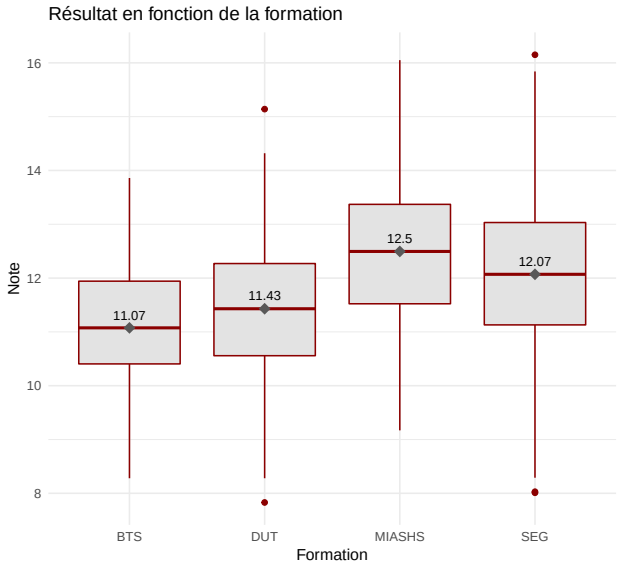


TABLE 4 – Taux de réussite de la série et de la mention au bac

	Série		Mention			
	ES	**S**	**AB**	**B**	**P**	**TB**
non	0.4510763	0.5420259	0.49	0.2157895	0.5785627	0.1794872
oui	0.5489237	0.4579741	0.51	0.7842105	0.4214373	0.8205128

Le fait d'avoit fait un bac S ou ES ne semble pas impacter les note. Le taux de réussite est différent de moins de 10% entre les deux filières. Les mentions bien et très bien au bac réussissent meieux que les mentions AB et passable. En effet, nous sommes aux alentours de 80% pour les mentions bien et très bien et 45-50% pour les mentions AB et Passable.



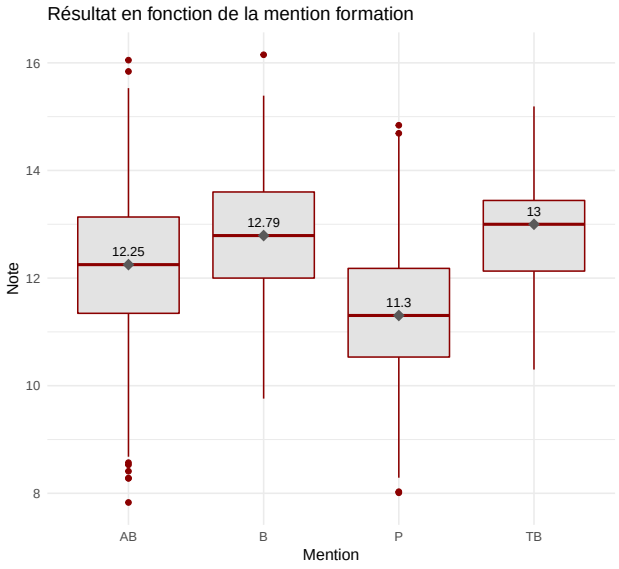


Table 5 – Taux de réussite en fonction de la formation et mention obtenue

	Formation				Mention			
	BTS	**DUT**	**MIASHS**	**SEG**	**AB**	**B**	**P**	**TB**
non	0.7608696	0.6798561	0.347032	0.4726277	0.4159613	0.2491803	0.694517	0.2307692
oui	0.2391304	0.3201439	0.652968	0.5273723	0.5840387	0.7508197	0.305483	0.7692308

Enfin, la formation suivie a un impact important, les formations telles que BTS et DUT réussissent beaucoup moins bien que les formations tels que MIASHS et SEG. De plus, la mention semble joué un rôle surtout pour la mention passable qui à un taux faible par rapport aux 3 autres. La mention TB et B ont quasiment le même taux de réussite.

La direction du département souhaiterait publier ces statistiques tout simplement pour déterminer qu'elles sont les résultats obtenus en fonction des charactéristiques des candidats. Cette publication sera informative pour les futurs candidats. Les conséquences de cette publication seraient que des élèves ayant des caractéristiques similaires aux personnes ayant eu de mauvaises notes, ne vont pas vouloir s'inscrire et tenter le concours. Cela pourrait donc démotiver des candidats potentiels.

2 Question 2

Pour cette question, nous allons réaliser un modèle linéaire avec les notes en variable à expliquer. Nous ne gardons pas la variable admission car elle est formé à partir de la variable note (corrélée à 100%) ainsi que la variable année (une regréssion préliminaire a été effectuée afin de s'assurer de la non-significativité).

Si on observe les coefficients (voir annexe), toutes les modalités de la variable retard sont significatives et ont un coefficient négatif, c'est-à-dire que plus l'année de retard est important, moins la note sera bonne. Les coefficients associés aux variables `formation` et `mention obtenue` sont également sigificatifs. Pour la variable `formation`, on remarque que les coefficients sont tous positifs, ce qui signifie que le fait de faire une formation différente de BTS (catégorie de référence) augmentera la note. De plus, seule la mention passable du bac est très significative avec un coefficient négatif, donc si l'individu prend cette modalité, sa note va baisser.

Analyse d'un coefficient : Par exemple pour la variable `retard`, la catégorie de référence est -1, ce qui signifie qu'un étudiant ayant un retard de 0 aura une baisse moyenne de la note finale de 0.58 par rapport à un élève en avance de 1 an. Un élève en retard de 3 ans aura une baisse moyenne significative de 2.17 sur sa note finale. On peut également voir que les variables `annee` et `nationalité` ont des coefficients non significatifs. Toutefois, ces variables ont quand même un intérêt statistique, sans qu'elles soient significatives, elles sont intéressantes à observer pour l'étude. On les gardera donc pour le futur. On peut donc faire le lien avec notre analyse statistique réalisée précèdemment ou l'on retrouve en variables ayant un coefficient significatif, les variables discriminantes et en variables ayant un coefficient non significatif, les variables qui semblaient non discriminantes dans notre analyse.

3 Question 3

Dans cette question, nous allons réaliser un modèle à probabilité linéaire. Pour cela, nous utilisons la même méthode que pour une regression linéaire simple mais on met en variable à expliquer, une variable dichotomique (ici l'`admissibilite`). Dans un modèle à probabilité linéaire, Y_i suit une loi de bernoulli. On a donc

$$E(Y_i) = 0(1 - P_i) + 1(P_i) = P_i$$

avec

$$E(Y_i|X_i) = B_0 + \sum B_k X_i = P_i$$

et doit être compris entre 0 et 1. De plus, dans ce type de modèle, nous allons chercher à estimer la probabilité associé à l'évenement P_i=1. Si $\hat{P}_i < 0.5$, on lui attribuera la valeur 0, si $\hat{P}_i > 0.5$, la valeur 1.

Voici la sortie R du test d'hétéroscédasticité :

```
##
##  studentized Breusch-Pagan test
##
## data:  model_lpm
## BP = 92.465, df = 16, p-value = 8.774e-13
```

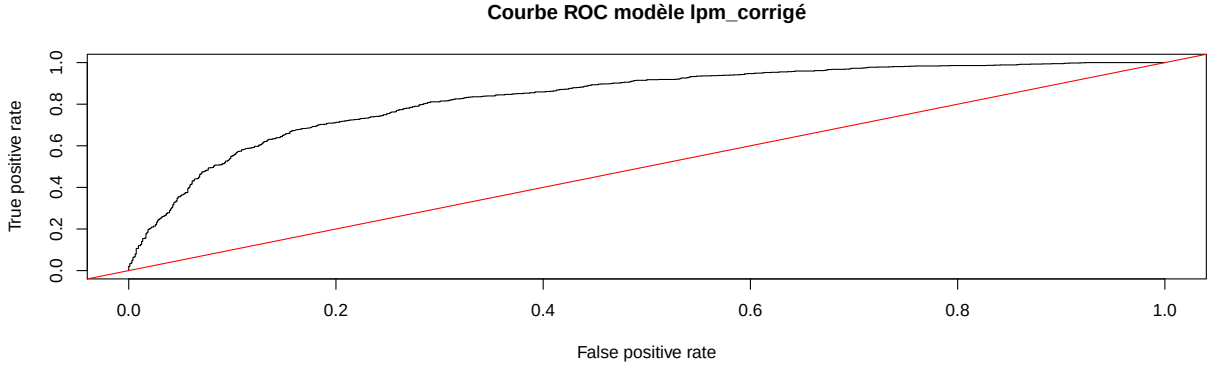
La statistique du test de Breush-Pagan est de 92.465, et la p-value associée est très faible (largement inférieur à 0.05). La p-value étant < 0.05, nous rejetons l'hypothèse H_O de présence d'homoscédasticité et donc on est en présence d'hétéroscédasticité. En effet, même si $E(\epsilon_i) = 0$, la distribution de l'erreur suivant une loi de bernoulli, nous avons $var(\epsilon_i) = P_i(1 - P_i)$ et dépend donc de X, ce qui cause l'hétéroscédasticité. Afin de résoudre ce problème, nous allons appliquer une pondération au modèle en utilisant la valeur prédite $\hat{Y}_i$. On aura $w_i = \hat{Y}_i * (1 - \hat{Y}_i)$ comme poids. Enfin,le modèle utilisera weights $= \frac{1}{\sqrt{w_i}}$ dans le modèle. Toutefois, cela pose un problème pour les valeurs prédites de probabilité < 0 et supérieur à 1 (ce qui montre la limite du modèle à probabilité linéaire). Nous allons donc affecter un valeur très proche de 0 si $\hat{Y}_i < 0$ et une valeur proche de 1 si $\hat{Y}_i > 1$.

Nous pouvons maintenant comparer les deux modèles (dont les résultats se trouvent en annexe) : On peut déja remarquer que le $R2$ a été amélioré pour le modèle corrigé, passant de 0.335 à 0.506. En ce qui concerne les écarts-types, ils ont tous été diminués dans le modèle corrigé et la valeur des coefficents est quasiment similaire entre les deux modèles. On remarque comme à la question 2, que les variables `retard`, `formation`, `mention obtenue` sont significatives. On retrouve également la non significativité des variables `nationalité` et `année`.
Voici un exemple d'interprétation : le fait d'avoir retard = 0, on a un coefficient négatif de -0.13, c'est à dire qu'il y a 13% de chance de moins d'être admissible. Un retard de 3 ans, entraine une baisse de 62% de chance d'être admissible par rapport à un étudiant en avance de 1 an (catégorie de référence). Un autre coefficient important est le fait d'avoir suivi une formation MIASHS qui augmente les chances d'être admissible de 42% par rapport à quelqu'un ayant effectué un BTS.
Voici la matrice de confusion associée au modèle :

TABLE 6 – Matrice de confusion lpm

	Prévision	
	0	**1**
0	706	258
1	217	769

Nous avons un taux de bien prédit de 75.641%. On peut également voir que l'on prédit un peu mieux les personnes admises que les personnes non admises.
Enfin, voici la représentation de la courbe ROC



La ligne rouge est la première bissectrice. On a une AUC de 0.833.

Toutefois l'interprétation de ce modèle à probabilité linéaire pose problème puisque les estimations peuvent aller de $-\infty$ à $+\infty$. On va donc utiliser des modèles qui assure une loi de distribution entre 0 et 1 : les modèles Logit et Probit.

4 Question 4

Dans cette question, nous allons réaliser un modèle logit. Au contraire du modèle à probabilité linéaire, ce modèle va pouvoir se modéliser comme une forme en "S". De plus, la distribution de l'erreur dans ce modèle suit une loi logistique avec $E(\epsilon) = \mu$ et $V(\epsilon) = \frac{s^2\pi^2}{3}$, or le modèle logit ne peut être exprimé que si $E(\epsilon) = 0$ et $V(\epsilon) = \frac{\pi^2}{3}$.
On retrouve en annexe, les résultat de cette regression. Les coefficients nous donnent une idée sur la relation entre variables explicatives et variables à expliquer, toutefois ils ne peuvent pas être interprétés directement. Pour cela, nous devons passer par les odds ratios. Pour les calculés, on prend l'exponentielle du coefficient. Voici un tableau des odds ratios du modèle logit :

4.1 Odds ratios et IC à 95%

TABLE 7 – Rapports de chance

	OR	2.5 %	97.5 %
(Intercept)	2.4845619	1.0960612	5.8951489
sexeH	0.8752280	0.7026462	1.0894357
nationalitefrançais	0.8759428	0.6614420	1.1588340
retard0	0.4368326	0.2144867	0.8228495
retard1	0.2093921	0.1025083	0.3961070
retard2	0.1050677	0.0500282	0.2053519
retard3	0.0327057	0.0154365	0.0644167
'série de bac'S	0.9667663	0.7720240	1.2116918
'mention de bac'B	1.3071401	0.8462866	2.0522957
'mention de bac'P	0.6273606	0.4951181	0.7937264
'mention de bac'TB	1.7633435	0.7184125	4.9204062
'formation suivie'DUT	1.7709822	1.0711993	2.9716228
'formation suivie'MIASHS	8.9311242	5.5224035	14.7432065
'formation suivie'SEG	4.1675068	2.6882988	6.6017007
'mention obtenue'B	1.8239898	1.2929273	2.5920507
'mention obtenue'P	0.4106654	0.3226099	0.5219484
'mention obtenue'TB	2.3185632	1.1215201	5.1136284

Interpretation des odds ratios

Tout d'abord, il est utile de rappeler que :

- Si le odd ratio est entre 0 et 1, alors cela correspond à un coefficient négatif et donc une relation negative.
- Si le odd ratio est supérieur à 1, cela correspond à une coefficient positif et donc une relation positive.
- Si le odd ratio = 1, alors il n'y a pas de différence entre les modalités ou la variable explicative n'a pas d'effet.

Dans notre modèle, on remarque que les coefficients non significatifs ont des odds ratios proches de 1, ce qui est logique car plus le ratio est proche de 1, moins l'effet est important.

Les étudiants en retard de 3 ans ont 0.03 fois plus de chance d'être admissible que les élèves en avance de 1 an. Plus précisement, les élèves en avance de 1 an ont 33.333 fois plus de chance d'être admis qu'un élève en retard de 3 ans. Une autre variable importante est la `formation suivie`, on peut voir que les étudiants ayant suivi la formation MIASHS ont 8.93 fois plus de chance d'être admis qu'un étudiant ayant fait un BTS et 4.16 fois plus de chance pour un élève ayant suivi SEG. Enfin pour la variable `mention obtenue`, les étudiants ont 2,3 et 1.8 fois plus de chance de réussir pour une mention TB et B respectivement que un étudiant ayant eu une mention AB. Au contraire un étudiant ayant eu une mention AB aura 2.4390244 fois plus de chance d'être admis qu'un élève avec mention passable.

4.2 Pseudo_R2

Nous avons un pseudo $R2$ de 0.273.

4.3 Matrice de confusion

Voici la matrice de confusion du modèle logit

Table 8 – Matrice de confusion logit

	Prévision	
	0	**1**
0	717	247
1	228	758

4.4 Différents estimateurs :

Ratio de mal prédit = 24.359%

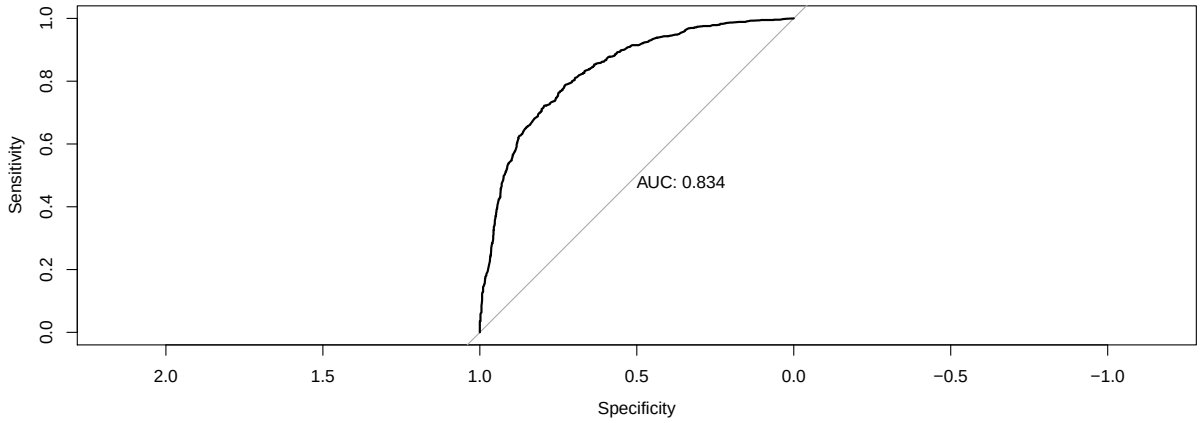
Ratio de bien prédit = 75.641%

Sensibilité : capacité à prédire correctement le succès = 76.876 %.

Spécificité : capacité à predire correctement l'échec = 74.378%

4.5 Courbe ROC

La courbe ROC est une représentation de l'arbitrage entre taux de faux positif et de vrais positif, c'est un arbitrage entre sensibilité et spécificité. L'aire sous la courbe ROC (AUC) mesure le pouvoir prédictif du modéle.



L'AUC du modèle logit est de 0.834.

5 Question 5

Nous allons maintenant estimer un modèle probit et le comparer à nos autres modèles afin de choisir le meilleur modèle final pour nos prédictions. Les erreurs du modèle probit suivent une loi normale comparé au modèle logit ou les erreurs suivent une loi logistique. Le modèle probit peut être mis en place seulement si $E(\epsilon) = 0$ et $V(\epsilon) = 1$. Il faut également noter que nous ne pouvons pas comparer les coefficients du modèle logit et probit puisque les normalisations ne sont pas les mêmes. Pour comparer les deux modèles, on utilise le facteur $\frac{\pi}{\sqrt{3}}$, ce qui signifie que $\beta_{logit} \sim \frac{\pi}{\sqrt{3}} * \beta_{probit}$. (Notons que le signe du coefficient restera le même entre logit et probit).

```
##
## ================================================================================================
##                                              Dependent variable:
##                          ------------------------------------------------------------------------
##                                                admissibilite                  `note épreuves écrites`
##                          logistic   probit             OLS                          OLS
##                             (1)       (2)              (3)                          (4)
## ------------------------------------------------------------------------------------------------
## annee-5                                0.080            0.006
##                                       (0.117)          (0.030)
##
## annee-4                               -0.074           -0.041
##                                       (0.116)          (0.027)
##
## annee-3                               -0.086           -0.032
##                                       (0.113)          (0.026)
##
## annee-2                                0.045           0.046*
##                                       (0.114)          (0.026)
##
## annee-1                               -0.088           -0.038
##                                       (0.113)          (0.027)
##
## sexeH                    -0.133       -0.078           -0.010                        0.004
##                          (0.112)      (0.065)          (0.017)                      (0.045)
##
## nationalitefrançais      -0.132       -0.077           -0.021                       -0.030
##                          (0.143)      (0.083)          (0.023)                      (0.057)
##
## retard0                  -0.828**     -0.481**         -0.131***                    -0.580***
##                          (0.340)      (0.187)          (0.044)                      (0.117)
##
## retard1                  -1.564***    -0.922***        -0.283***                    -1.012***
##                          (0.342)      (0.189)          (0.046)                      (0.120)
##
## retard2                  -2.253***    -1.351***        -0.446***                    -1.505***
##                          (0.358)      (0.199)          (0.050)                      (0.127)
##
## retard3                  -3.420***    -2.026***        -0.609***                    -2.174***
##                          (0.362)      (0.199)          (0.045)                      (0.123)
##
## `série de bac`S          -0.034       -0.029           -0.020                       -0.029
##                          (0.115)      (0.067)          (0.018)                      (0.046)
##
```

```
## `mention de bac`B          0.268     0.145            0.038                    0.160*
##                           (0.226)   (0.127)          (0.032)                  (0.082)
##
## `mention de bac`P        -0.466*** -0.276***        -0.064***                -0.261***
##                           (0.120)   (0.071)          (0.019)                  (0.049)
##
## `mention de bac`TB         0.567     0.325            0.085                    0.106
##                           (0.486)   (0.270)          (0.058)                  (0.165)
##
## `formation suivie`DUT     0.572**   0.341**          0.128***                 0.408***
##                           (0.260)   (0.151)          (0.040)                  (0.104)
##
## `formation suivie`MIASHS 2.190***  1.303***          0.424***                 1.379***
##                           (0.250)   (0.145)          (0.039)                  (0.100)
##
## `formation suivie`SEG    1.427***  0.864***          0.300***                 0.964***
##                           (0.229)   (0.133)          (0.035)                  (0.092)
##
## `mention obtenue`B       0.601***  0.349***          0.097***                 0.389***
##                           (0.177)   (0.102)          (0.028)                  (0.070)
##
## `mention obtenue`P       -0.890*** -0.523***        -0.169***                -0.515***
##                           (0.123)   (0.072)          (0.021)                  (0.051)
##
## `mention obtenue`TB       0.841**   0.496**          0.133**                  0.442***
##                           (0.385)   (0.219)          (0.056)                  (0.144)
##
## Constant                  0.910**   0.551**          0.640***                 12.442***
##                           (0.427)   (0.250)          (0.061)                  (0.156)
##
## ------------------------------------------------------------------------------------------
## Observations               1,950     1,950            1,950                    1,950
## R2                                                    0.506                    0.468
## Adjusted R2                                           0.501                    0.463
## Log Likelihood           -982.799  -980.279
## Akaike Inf. Crit.        1,999.597 2,004.558
## Residual Std. Error                              0.677 (df = 1928)        0.989 (df = 1933)
## F Statistic                                  94.094*** (df = 21; 1928) 106.123*** (df = 16; 1933)
## ==========================================================================================
## Note:                                                          *p<0.1; **p<0.05; ***p<0.01
```

Voici la matrice de confusion associée au modèle probit et le taux de bien prédit.

Table 9 – Matrice de confusion probit

	Prévision	
	0	**1**
0	712	252
1	234	752

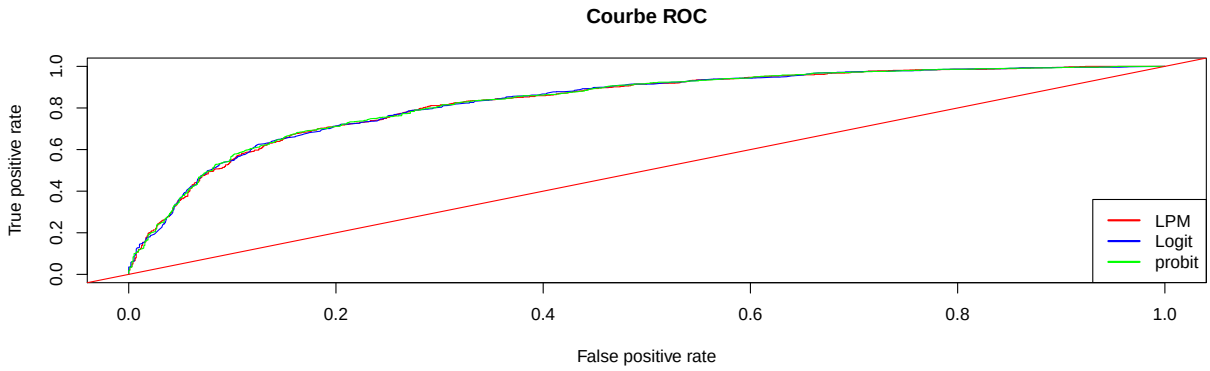
Le taux de bien prédit est de 75.077%.

Le taux de mal prédit est de 24.923%.

Sensibilité : capacité à prédire correctement le succès = 76.268 %.

Spécificité : capacité à predire correctement l'échec = 73.859%

5.1 Comparaison graphique :



On peut remarquer sur ce graphique que l'on peut à peine différencier les différents modèles. Cela denote la similitude de nos modèles et leurs pouvoirs prédictifs quasiment similaires.

L'AUC du modèle probit est de 0.835.

5.2 Comparaison des résultats

Tout d'abord, au vu des résultat, nous pouvons déjà mettre de coté le modèle économétrique de regression linéaire simple. En effet, celui-ci prédit la note d'un étudiant, or nous voulons prédire si l'étudiant va être admissible ou non, donc un modèle binaire. De plus, le $R2$ associé (0.468) est plus faible que pour un modèle à probabilité linéaire corrigé.

En ce qui concerne le modèle à probabilité linéaire, nous pouvons également le mettre de coté. Malgrès sa facilité dans sa création, il pose des problèmes, notamment celui de l'hétéroscédasticité, mais également le fait que le Y_i prédit peut être en dehors de 0-1. Enfin, un dernier problème sera le fait que nous voulons obtenir une courbe plus en forme de "S" plutôt qu'une droite et les modèles probit et logit seront donc plus adaptés. Toutefois, il sera difficile de choisir un des deux modèles car ils sont très similaires en terme d'ajutement statistique. Les différences se font pour de grand échantillons, ici avec 2000 données, l'échantillon ne semble pas assez grand.

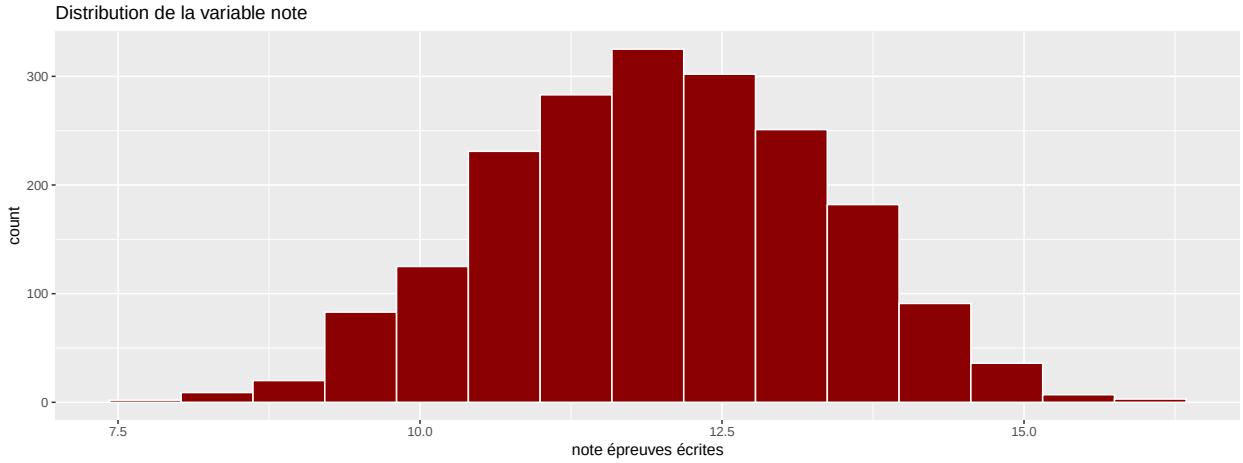
Finalement, pour pouvoir comparer le modèle probit au modèle logit, il nous faut déterminer le modèle qui à le meilleur taux de bonnes prédictions grâce aux matrices de confusion. Le ratio est :

$$\frac{\text{Nombres de bonnes prédictions}}{\text{Nombres d'observations}}$$

En ce basant sur ce ratio, nous avons un taux de bonnes prédictions de 75.641% pour le modèle logit et de 75.077% pour le modèle probit. Rappellons que le taux de bonnes prédictions pour le modèle à probabilité linéaire corrigé est de 75.641%. Nous avons donc les mêmes taux de réussite pour le modèle à probabilité linéaire et le modèle logit. Le meilleur modèle pour évaluer les chances de réussite pour un étudiant sera donc le modèle logit dans notre cas. (En ce basant sur ce qui a été vu précèdemment concernant les limites du modèle à probabilité linéaire).

6 Annexe

6.1 Question 1



La distribution des notes semblent suivre une distribution normale de d'espérance μ=12. Donc un taux d'amission général autour de 0,5.

6.2 Question 2

```
##
## ================================================
##                            Dependent variable:
##                        ------------------------
##                          `note épreuves écrites`
## ------------------------------------------------
## sexeH                              0.004
##                                   (0.045)
##
## nationalitefrançais               -0.030
##                                   (0.057)
##
## retard0                          -0.580***
##                                   (0.117)
##
## retard1                          -1.012***
##                                   (0.120)
##
## retard2                          -1.505***
##                                   (0.127)
##
## retard3                          -2.174***
##                                   (0.123)
##
## `série de bac`S                   -0.029
##                                   (0.046)
##
## `mention de bac`B                 0.160*
##                                   (0.082)
##
```

```
## `mention de bac`P                 -0.261***
##                                    (0.049)
##
## `mention de bac`TB                  0.106
##                                    (0.165)
##
## `formation suivie`DUT             0.408***
##                                    (0.104)
##
## `formation suivie`MIASHS          1.379***
##                                    (0.100)
##
## `formation suivie`SEG             0.964***
##                                    (0.092)
##
## `mention obtenue`B                0.389***
##                                    (0.070)
##
## `mention obtenue`P                -0.515***
##                                    (0.051)
##
## `mention obtenue`TB               0.442***
##                                    (0.144)
##
## Constant                          12.442***
##                                    (0.156)
##
## ------------------------------------------------
## Observations                         1,950
## R2                                   0.468
## Adjusted R2                          0.463
## Residual Std. Error           0.989 (df = 1933)
## F Statistic              106.123*** (df = 16; 1933)
## ================================================
## Note:                    *p<0.1; **p<0.05; ***p<0.01
```

6.3 Question 3

```
##
## ============================================================================
##                                        Dependent variable:
##                                ---------------------------------------------
##                                            admissibilite
##                                     (1)                      (2)
## ----------------------------------------------------------------------------
## sexeH                              -0.022                   -0.010
##                                    (0.019)                  (0.017)
##
## nationalitefrançais                -0.022                   -0.021
##                                    (0.024)                  (0.023)
##
## retard0                           -0.137***                -0.131***
##                                    (0.048)                  (0.044)
##
```

```
## retard1                          -0.288***              -0.283***
##                                   (0.050)                (0.046)
##
## retard2                          -0.438***              -0.446***
##                                   (0.053)                (0.050)
##
## retard3                          -0.637***              -0.609***
##                                   (0.051)                (0.045)
##
## `série de bac`S                   -0.010                 -0.020
##                                   (0.019)                (0.018)
##
## `mention de bac`B                  0.037                  0.038
##                                   (0.034)                (0.032)
##
## `mention de bac`P                -0.080***              -0.064***
##                                   (0.020)                (0.019)
##
## `mention de bac`TB                 0.077                  0.085
##                                   (0.068)                (0.058)
##
## `formation suivie`DUT            0.117***               0.128***
##                                   (0.043)                (0.040)
##
## `formation suivie`MIASHS         0.428***               0.424***
##                                   (0.041)                (0.039)
##
## `formation suivie`SEG            0.292***               0.300***
##                                   (0.038)                (0.035)
##
## `mention obtenue`B               0.109***               0.097***
##                                   (0.029)                (0.028)
##
## `mention obtenue`P               -0.174***              -0.169***
##                                   (0.021)                (0.021)
##
## `mention obtenue`TB               0.146**                0.133**
##                                   (0.060)                (0.056)
##
## annee-5                                                   0.006
##                                                          (0.030)
##
## annee-4                                                  -0.041
##                                                          (0.027)
##
## annee-3                                                  -0.032
##                                                          (0.026)
##
## annee-2                                                  0.046*
##                                                          (0.026)
##
## annee-1                                                  -0.038
##                                                          (0.027)
##
```

```
## Constant                         0.647***                  0.640***
##                                   (0.065)                   (0.061)
##
## ------------------------------------------------------------------------
## Observations                       1,950                     1,950
## R2                                 0.335                     0.506
## Adjusted R2                        0.330                     0.501
## Residual Std. Error          0.409 (df = 1933)         0.677 (df = 1928)
## F Statistic              60.862*** (df = 16; 1933) 94.094*** (df = 21; 1928)
## ========================================================================
## Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

6.4 Question 4

```
##
## Call:
## glm(formula = admissibilite ~ sexe + nationalite + retard + `série de bac` +
##     `mention de bac` + `formation suivie` + `mention obtenue`,
##     family = binomial(link = logit), data = data_logit)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2492  -0.7530   0.3887   0.7292   2.3669
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 0.9101     0.4271   2.131 0.033110 *
## sexeH                      -0.1333     0.1118  -1.192 0.233372
## nationalitefrançais        -0.1325     0.1430  -0.927 0.354178
## retard0                    -0.8282     0.3404  -2.433 0.014974 *
## retard1                    -1.5635     0.3422  -4.568 4.91e-06 ***
## retard2                    -2.2532     0.3579  -6.296 3.06e-10 ***
## retard3                    -3.4202     0.3622  -9.443  < 2e-16 ***
## `série de bac`S            -0.0338     0.1149  -0.294 0.768716
## `mention de bac`B           0.2678     0.2256   1.187 0.235103
## `mention de bac`P          -0.4662     0.1203  -3.874 0.000107 ***
## `mention de bac`TB          0.5672     0.4864   1.166 0.243571
## `formation suivie`DUT       0.5715     0.2598   2.200 0.027809 *
## `formation suivie`MIASHS    2.1895     0.2501   8.756  < 2e-16 ***
## `formation suivie`SEG       1.4273     0.2286   6.242 4.31e-10 ***
## `mention obtenue`B          0.6010     0.1773   3.391 0.000697 ***
## `mention obtenue`P         -0.8900     0.1227  -7.254 4.04e-13 ***
## `mention obtenue`TB         0.8409     0.3849   2.185 0.028909 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2703.0  on 1949  degrees of freedom
## Residual deviance: 1965.6  on 1933  degrees of freedom
## AIC: 1999.6
##
## Number of Fisher Scoring iterations: 4
```